**Accessing Our Data**

**Team member：**

**Kaiyang Lv**

**Ying Ma**

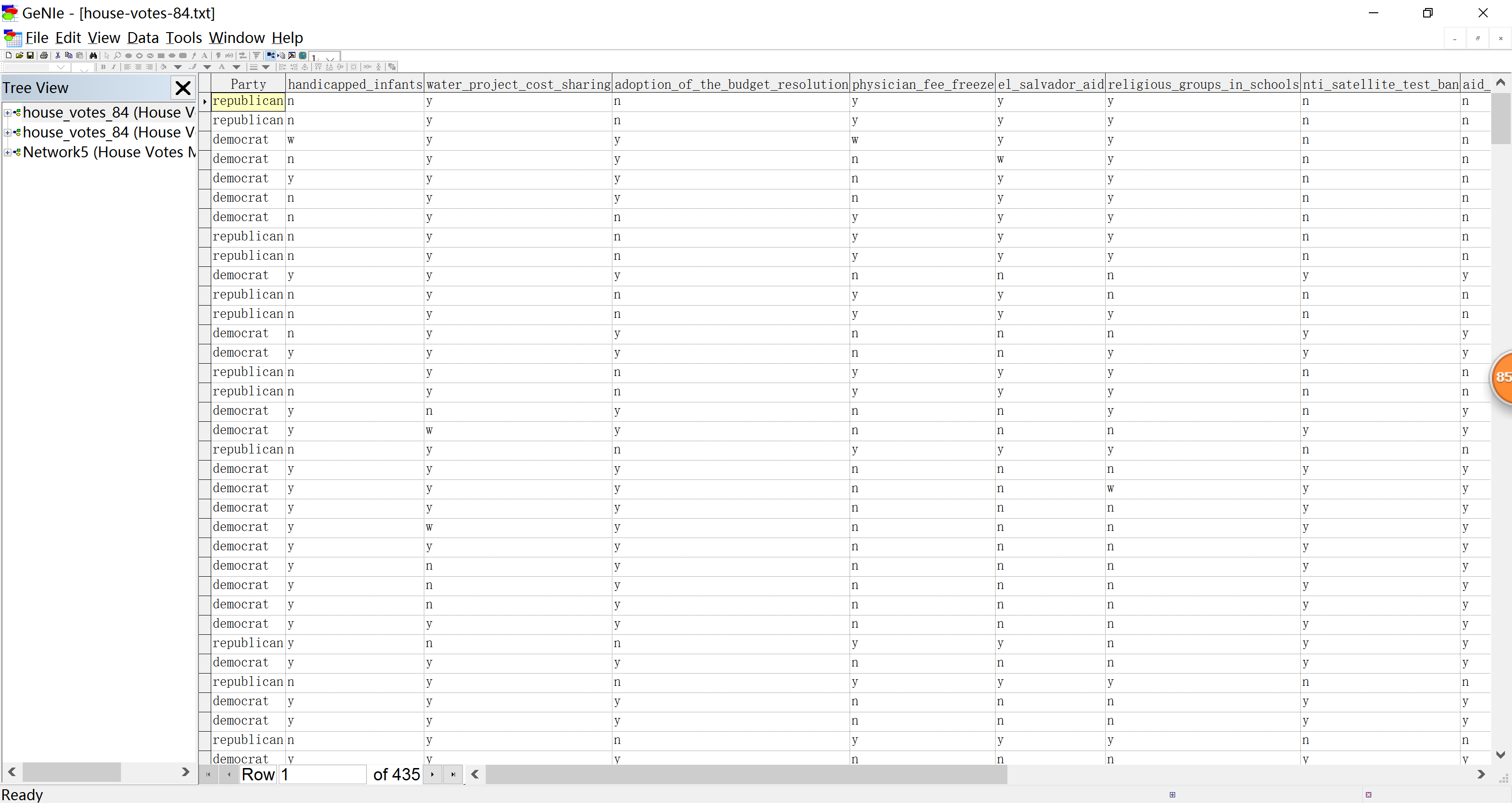
**Weiyi Tuo**

In this assignment we use GeNIe as our data analytic tool.

The first problem we encounter is how to access our data. Because we need to evaluate our three models, we need to set up our data for training and testing. After reading GeNIe’s documentation, we found the answer.



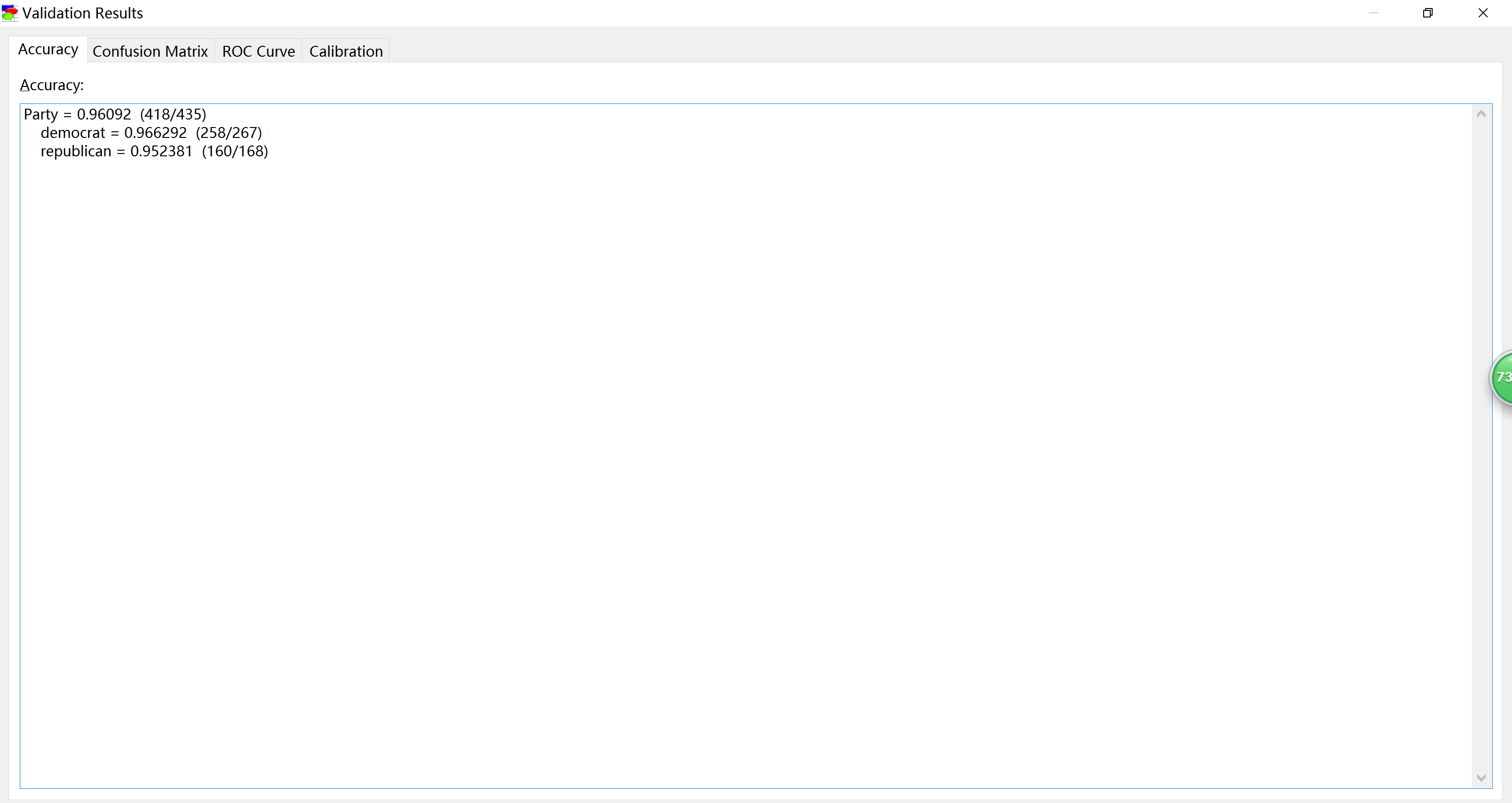
We input our data to the workplace.

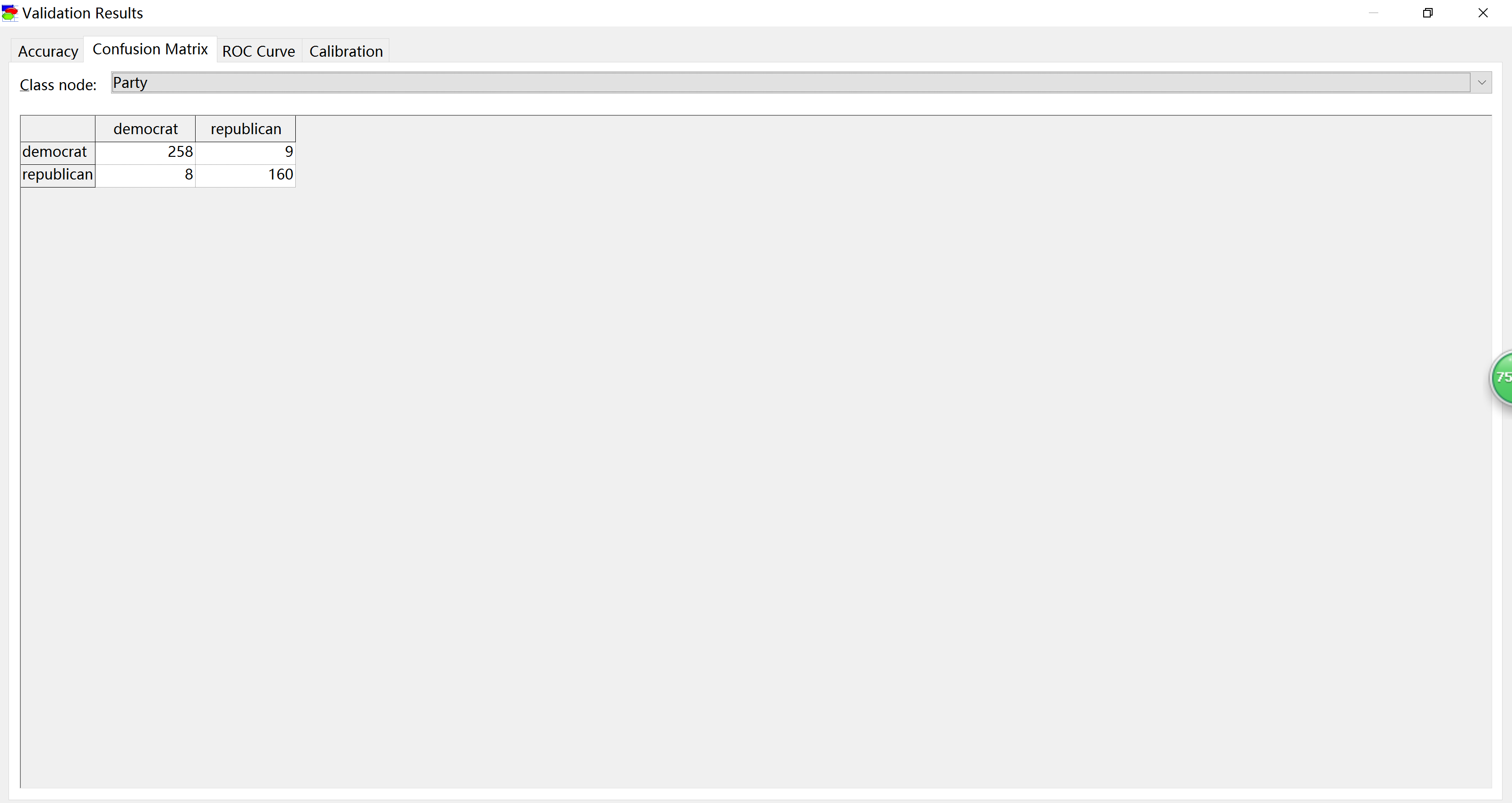


**Testing and Validation**

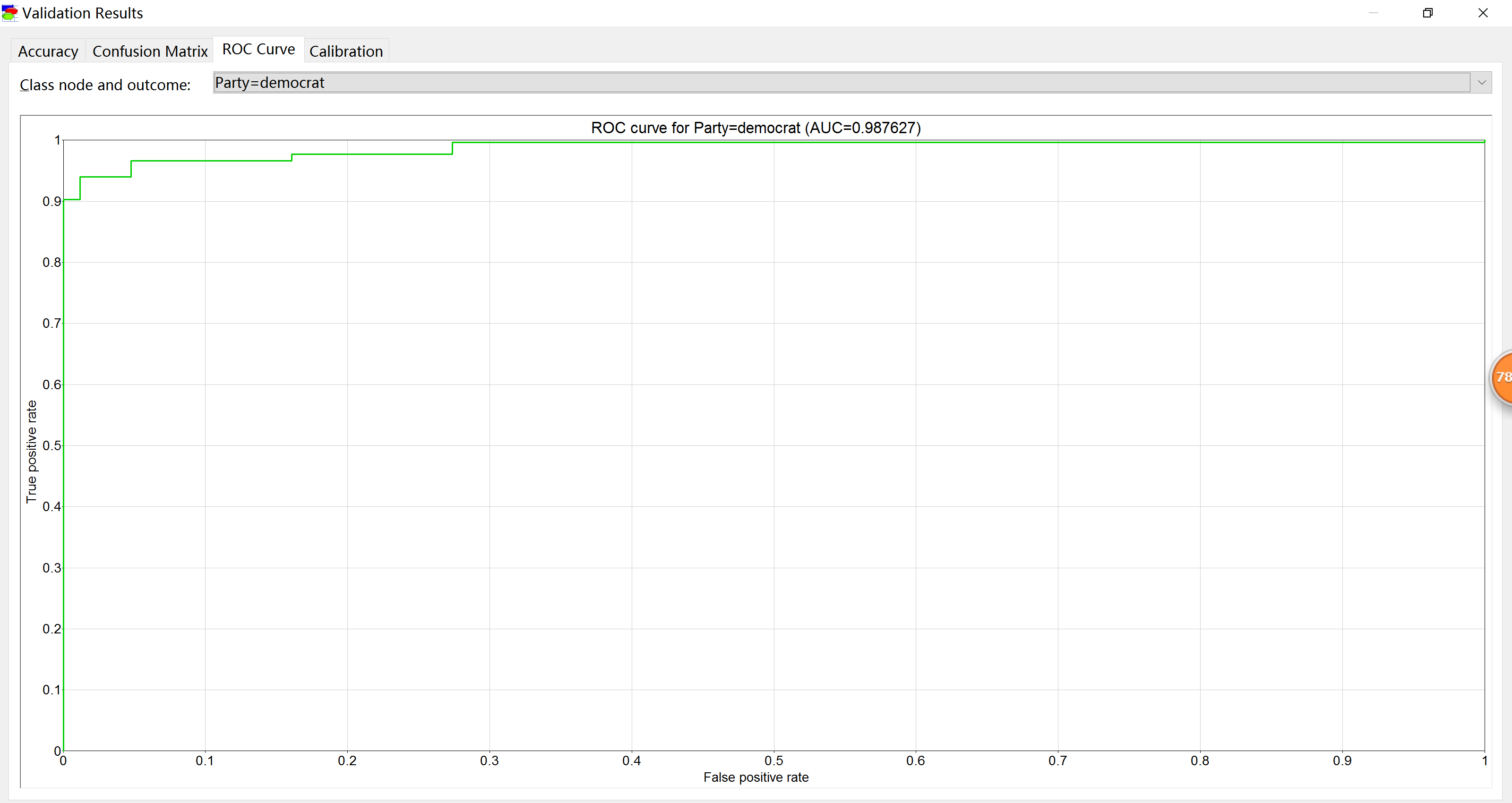
Input three models is quite simple work. We just dragged these three files with xdsl extension into workplace. Then, we are going to test the performance of these three models.

House Votes Manual model has 0.96092 of accuracy to predict the actual value.

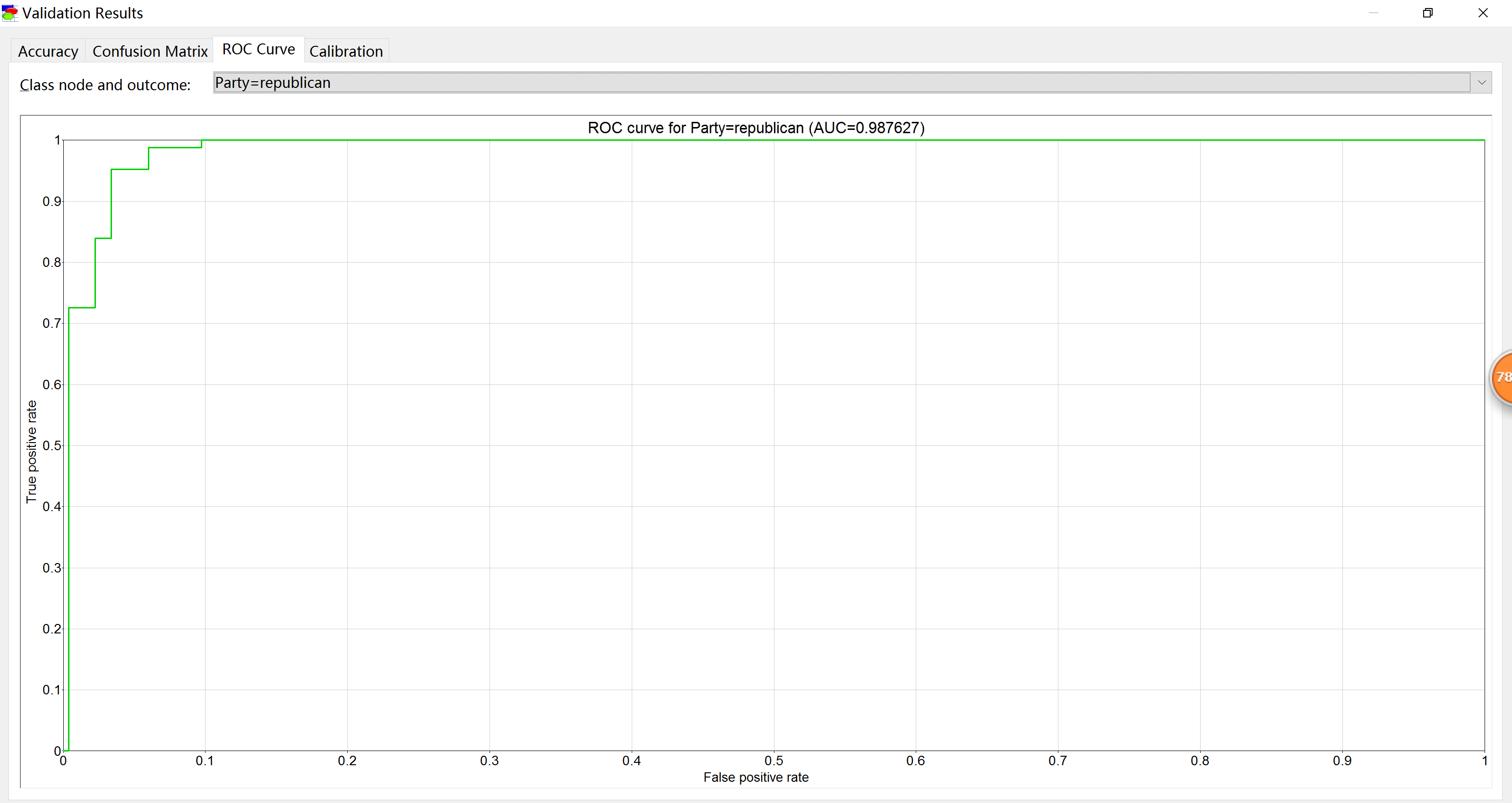




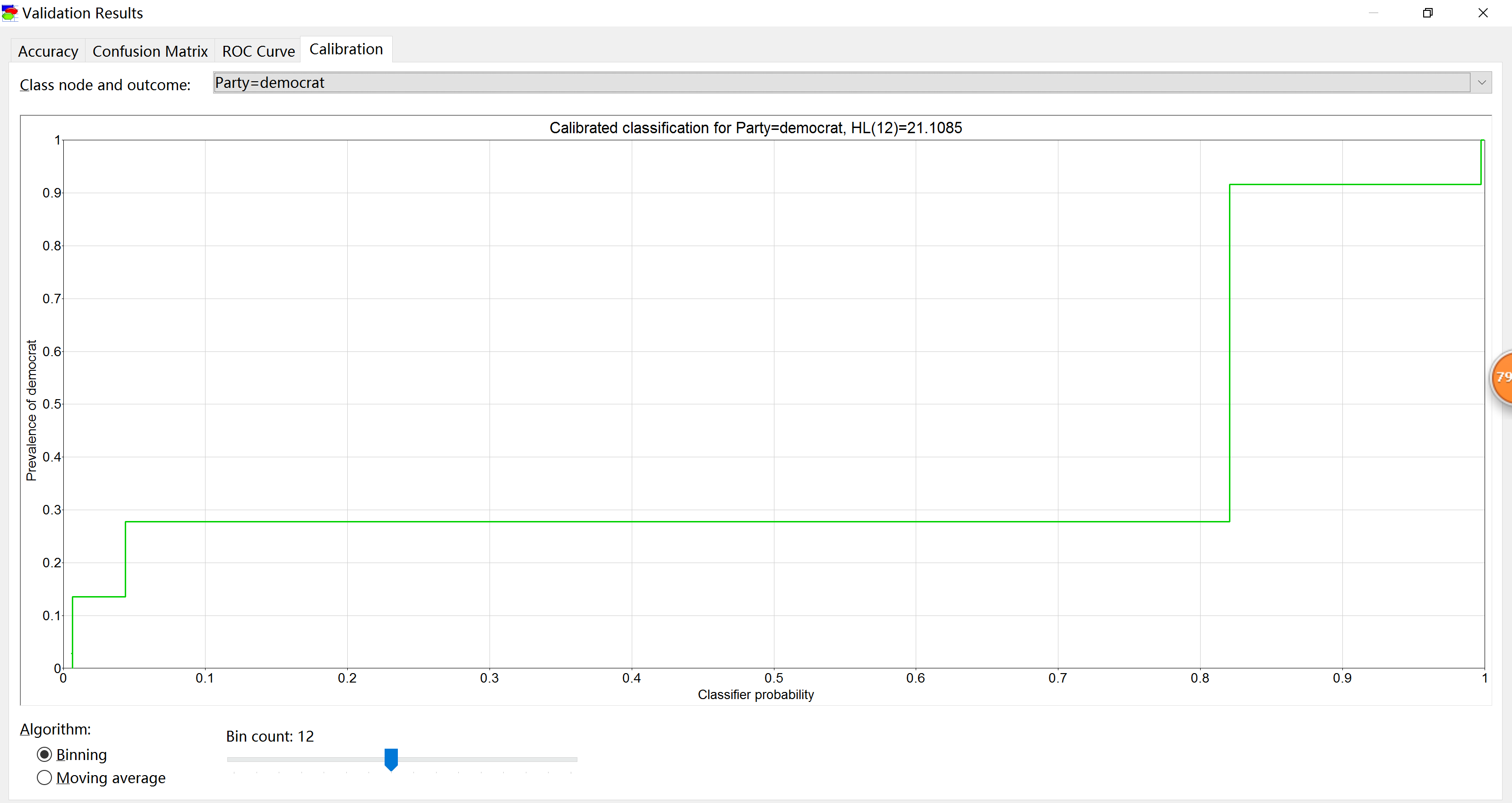
Then we generated ROC curve with Party = democrat, we can notice that AUC = 0.987627.



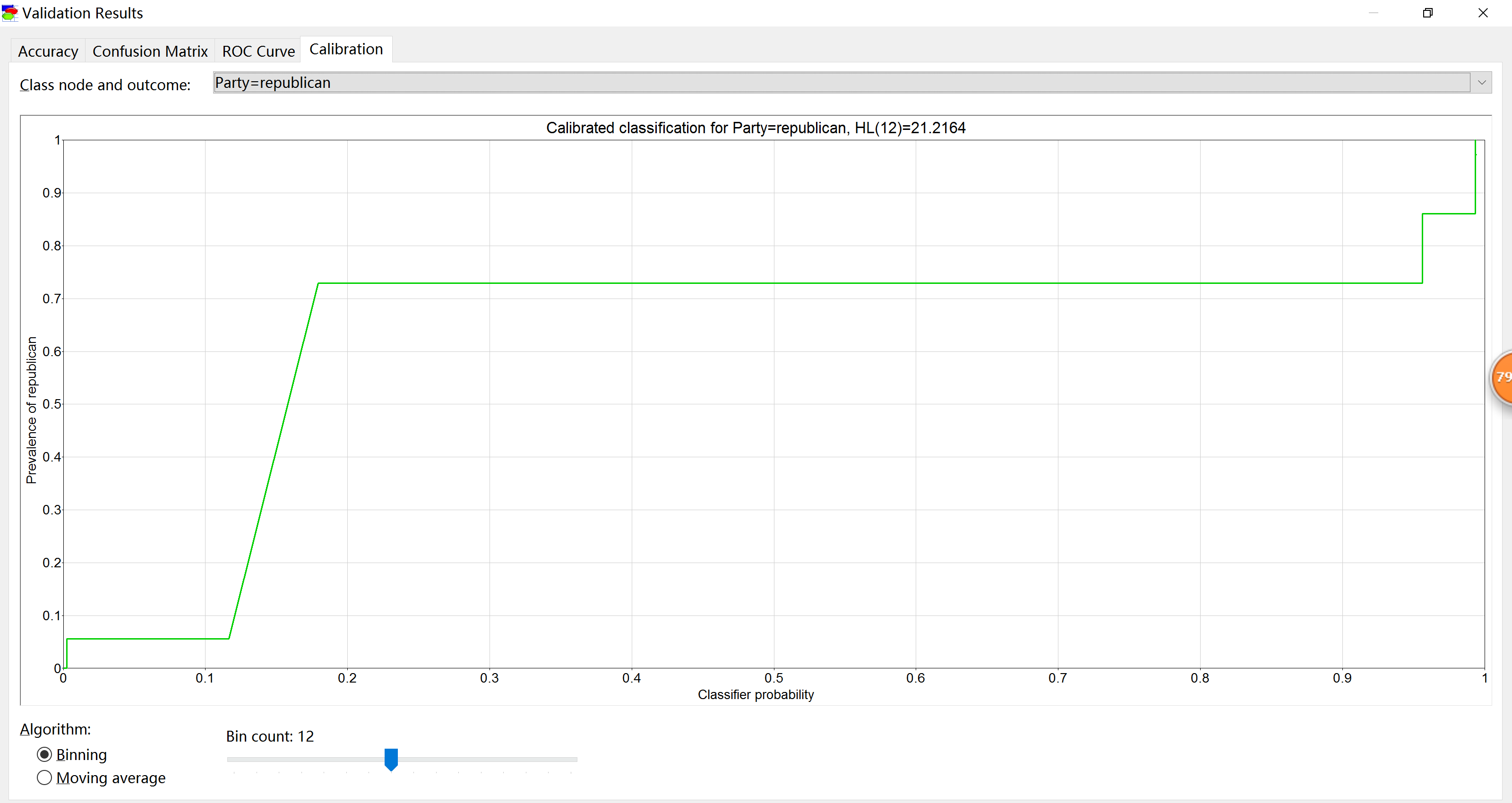
Also, here is ROC with Party = republican.



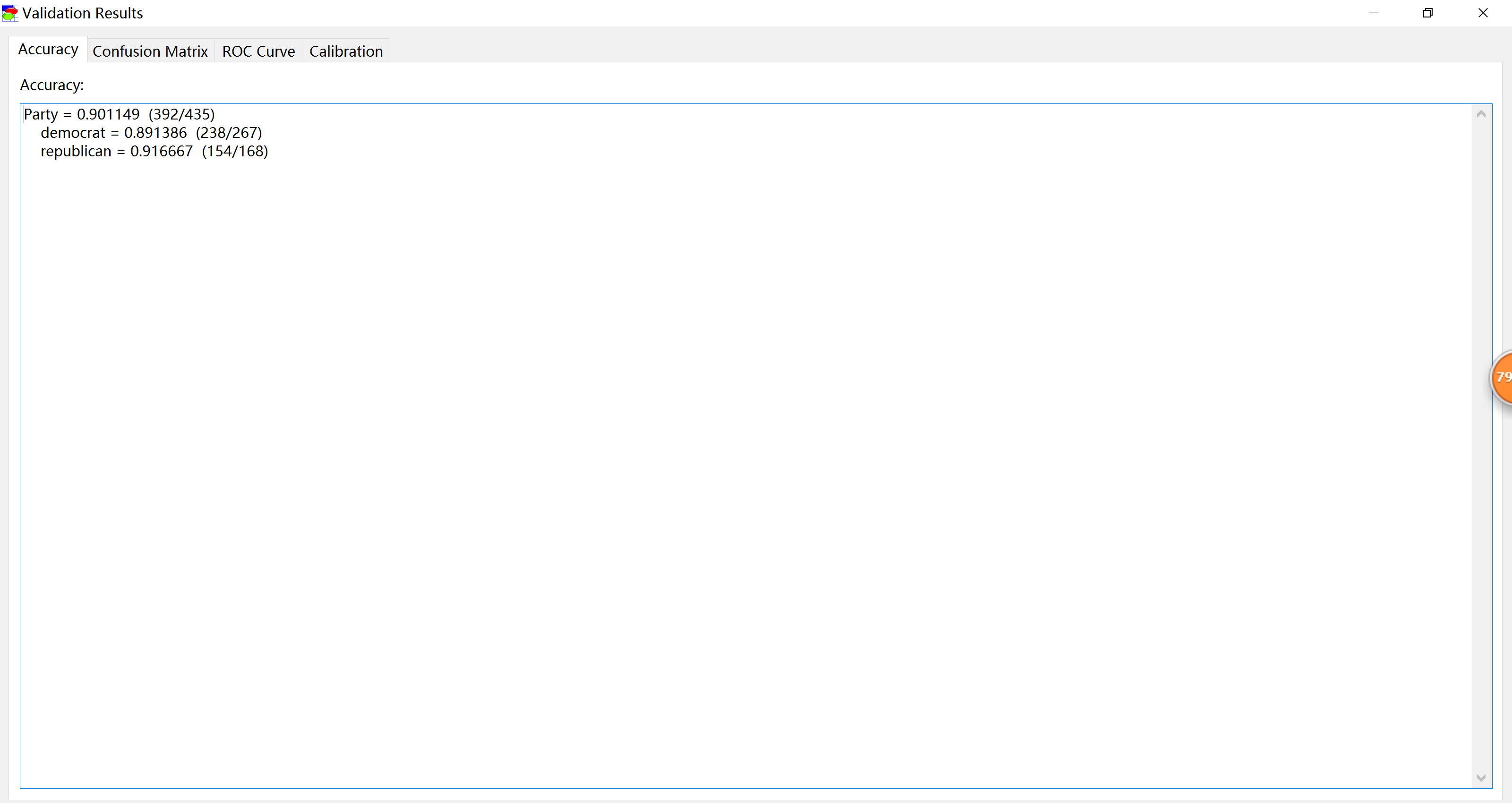
Next, we have democrat calibration for bin count = 12.



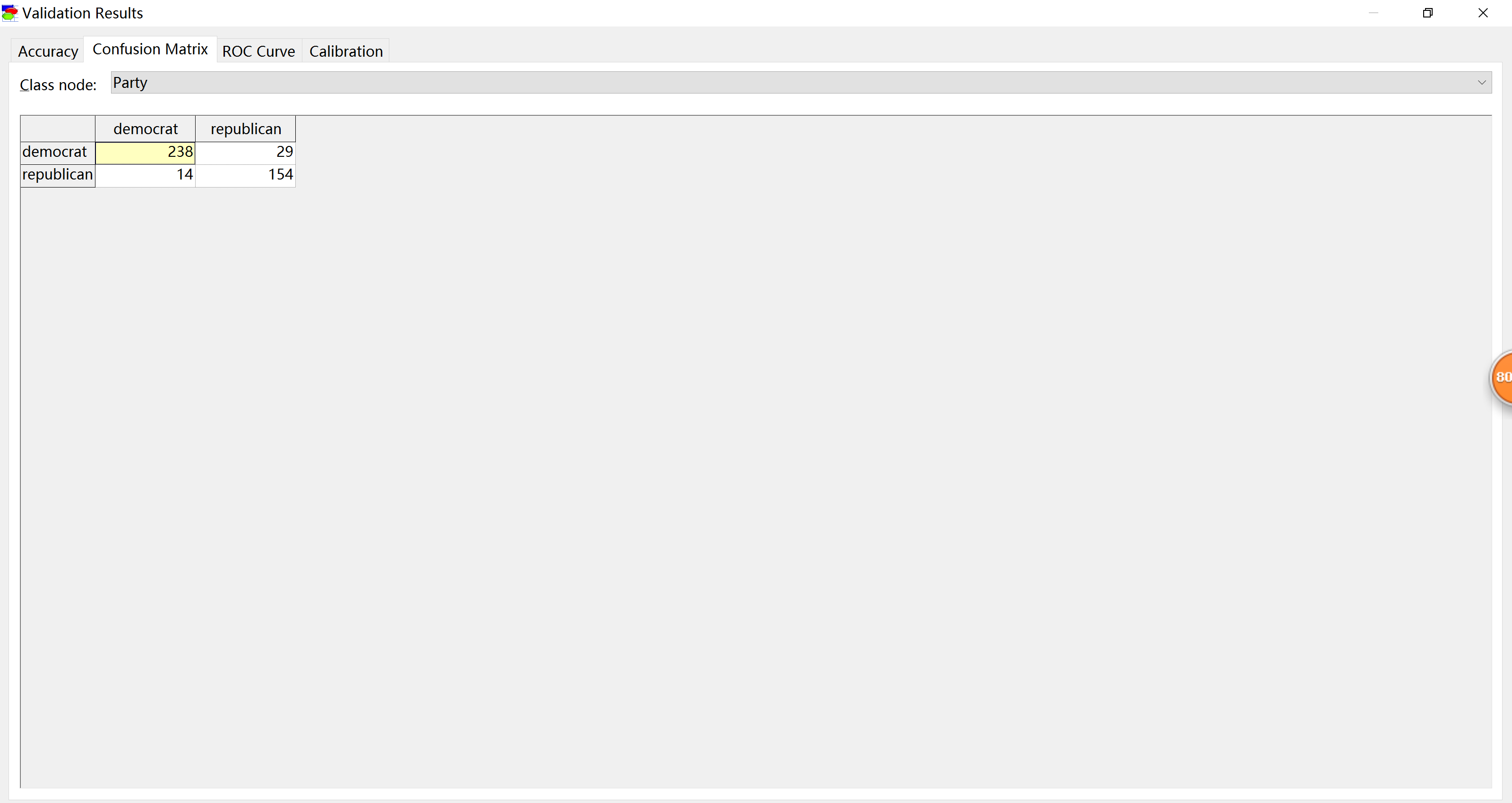
Plus, republican calibration for the same bin count.



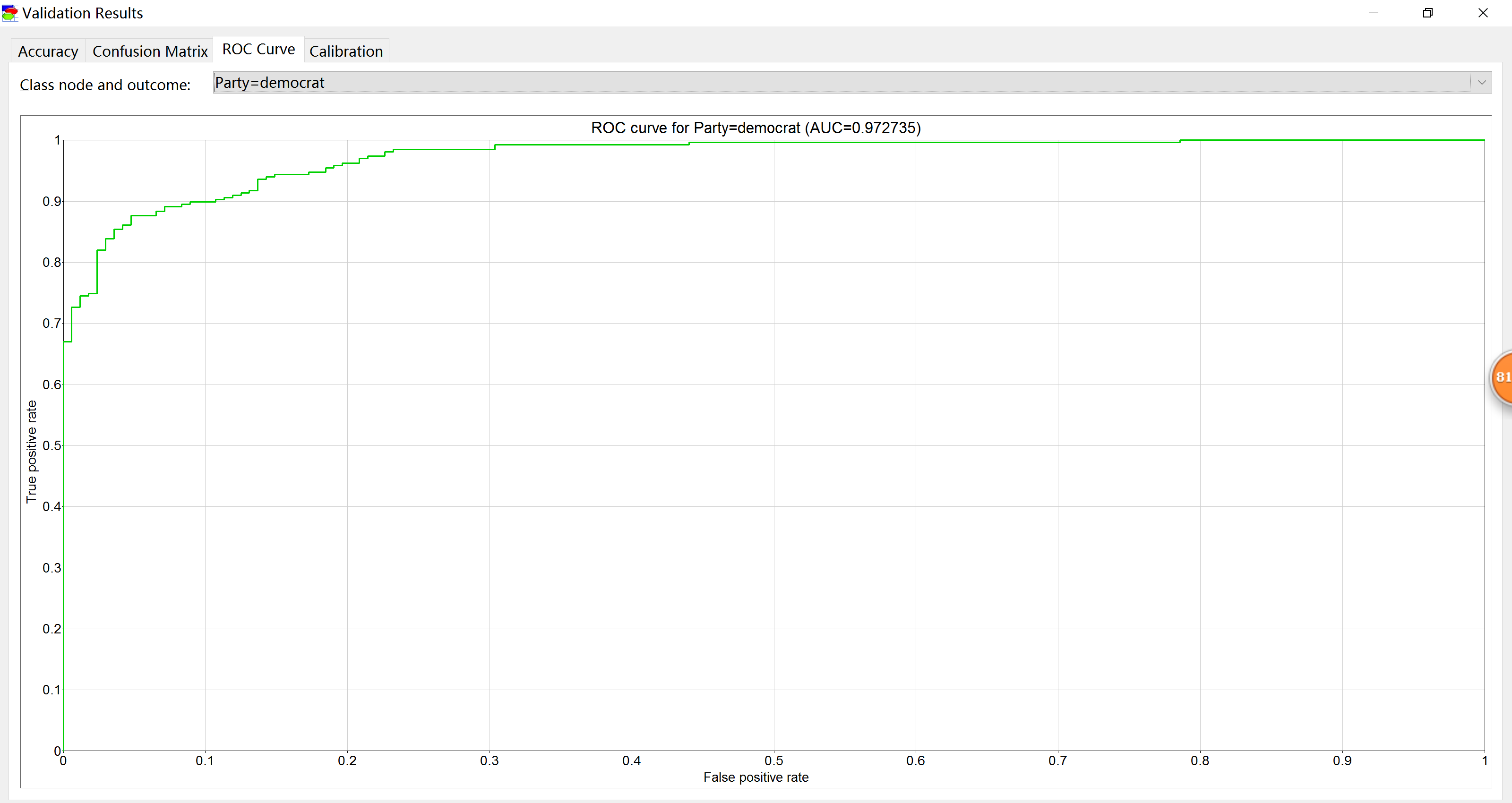
House Votes Naïve model has 0.901149 of accuracy to predict the actual value.



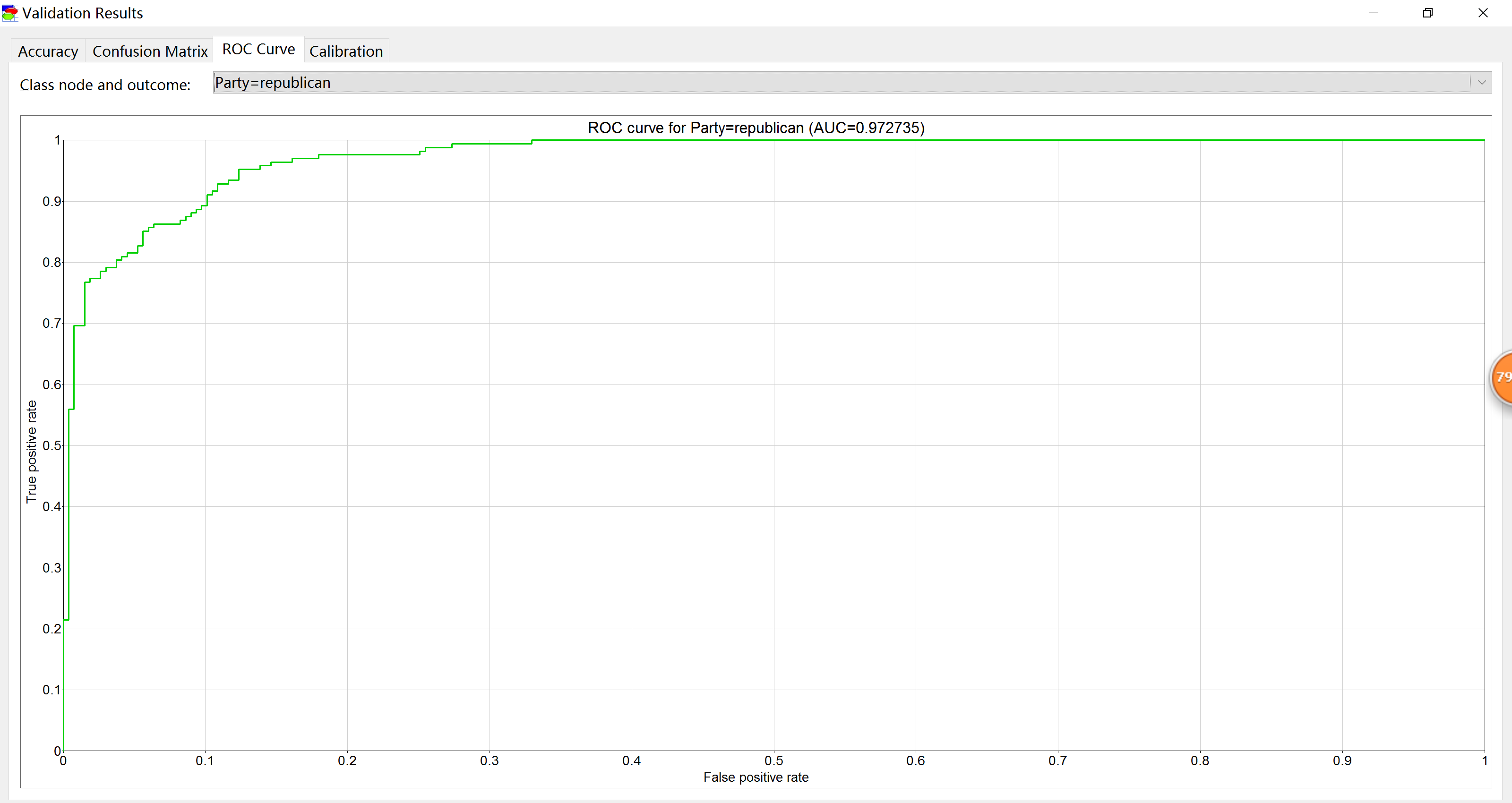
Then we got the confusion matrix



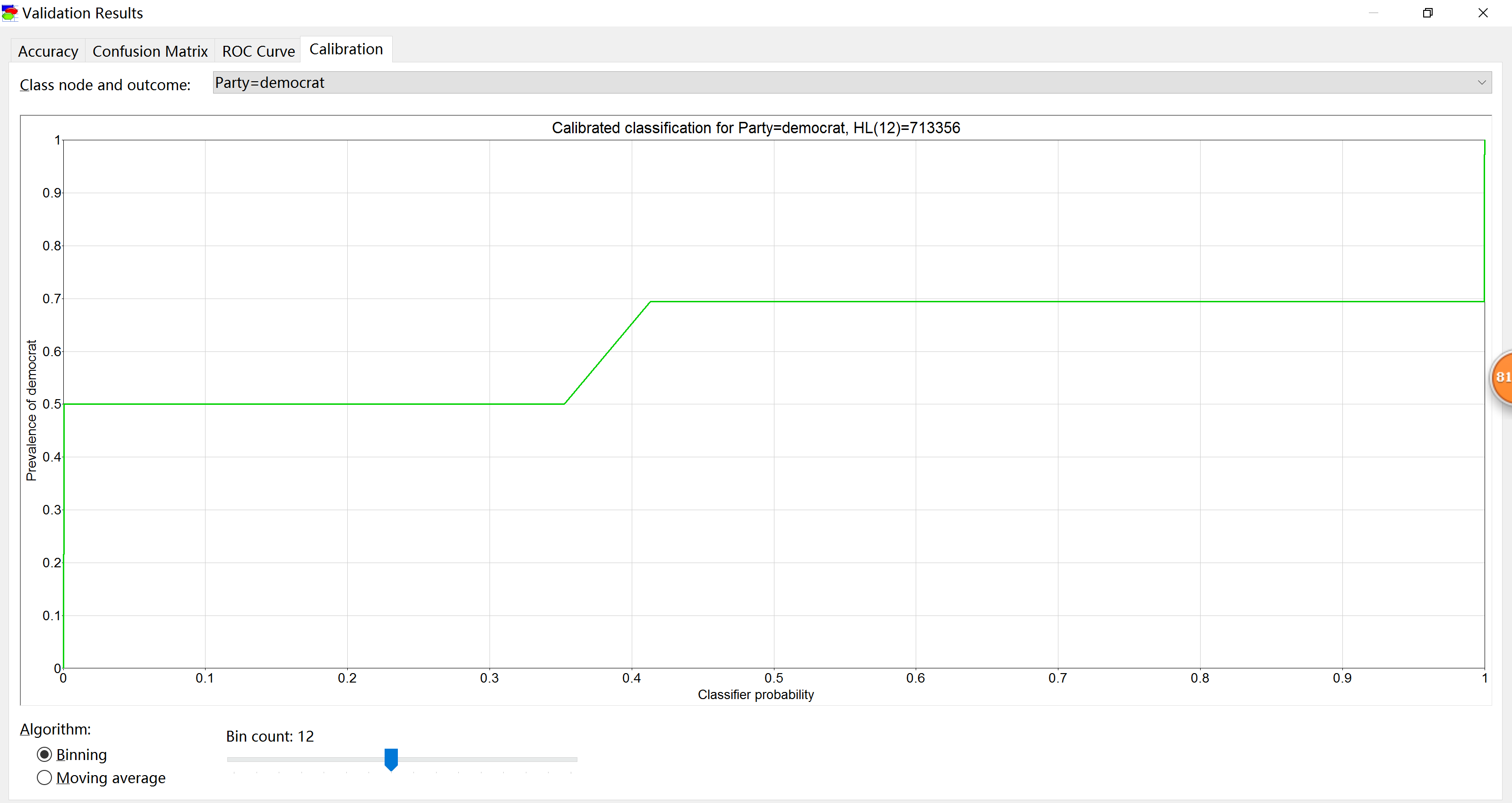
ROC curve with Party = democrat (AUC = 0.972735)

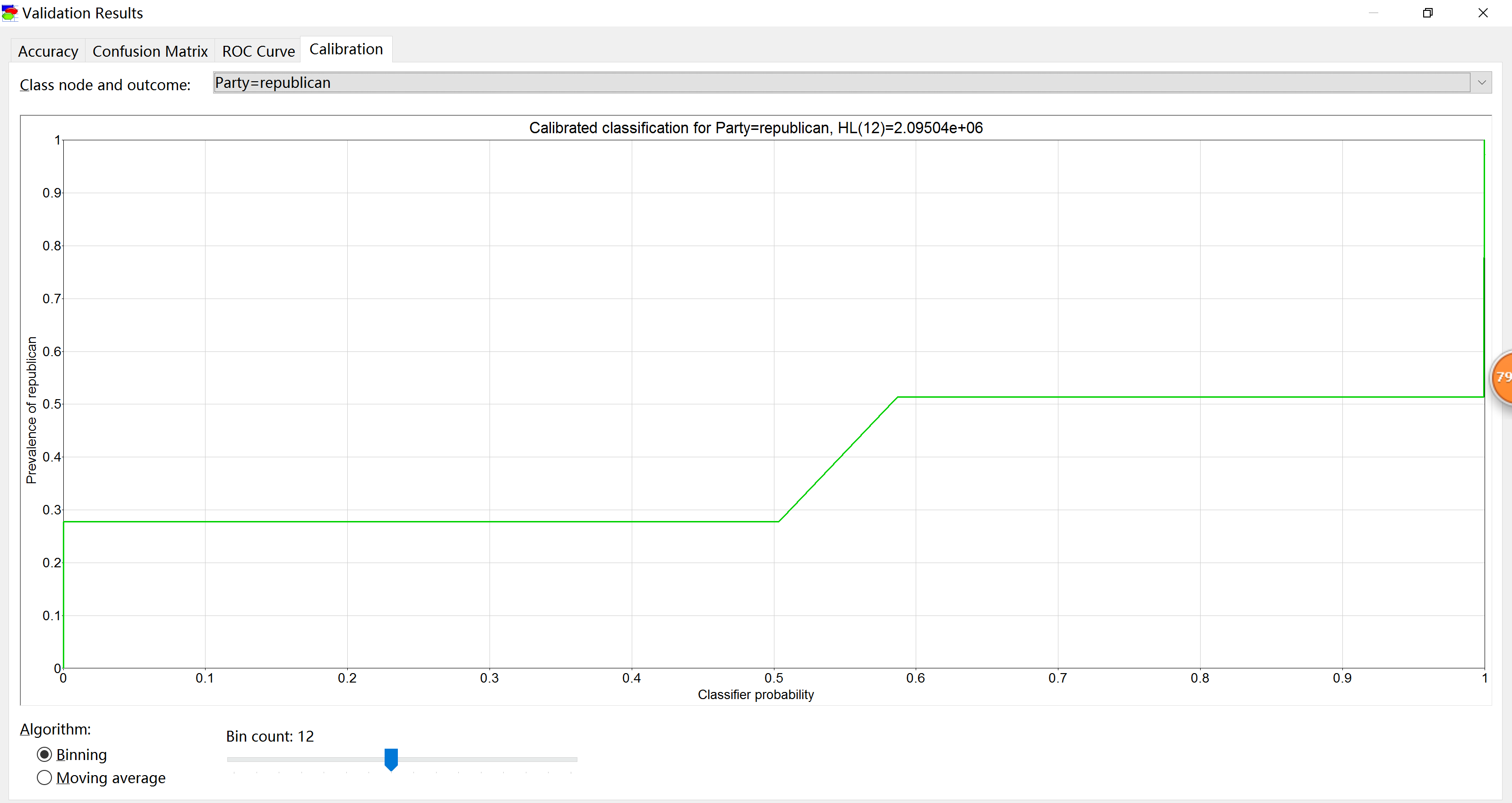


ROC curve with Party = republican

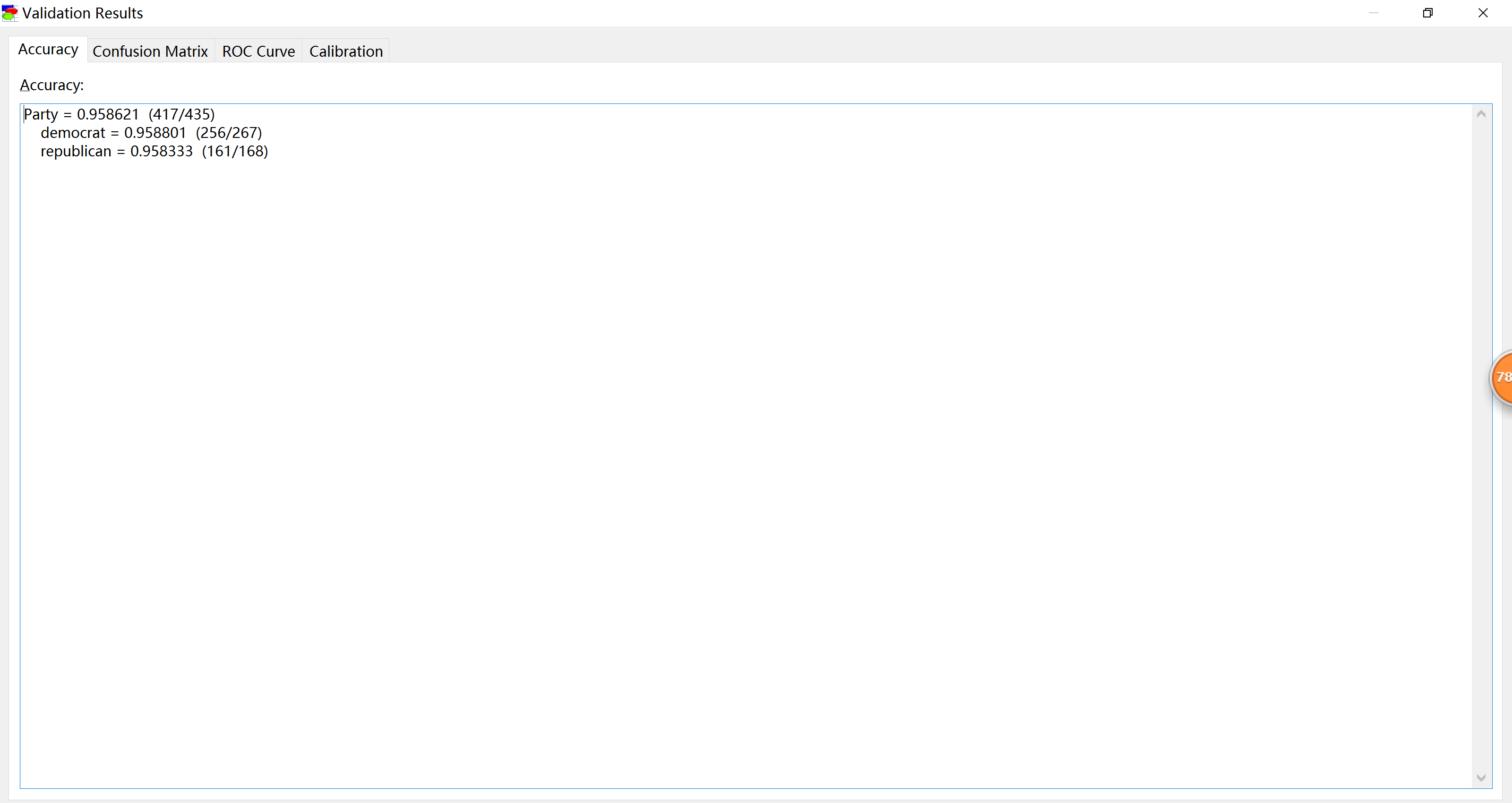


Since we did not change our selected bin count which is count = 12, we generated calibrations.

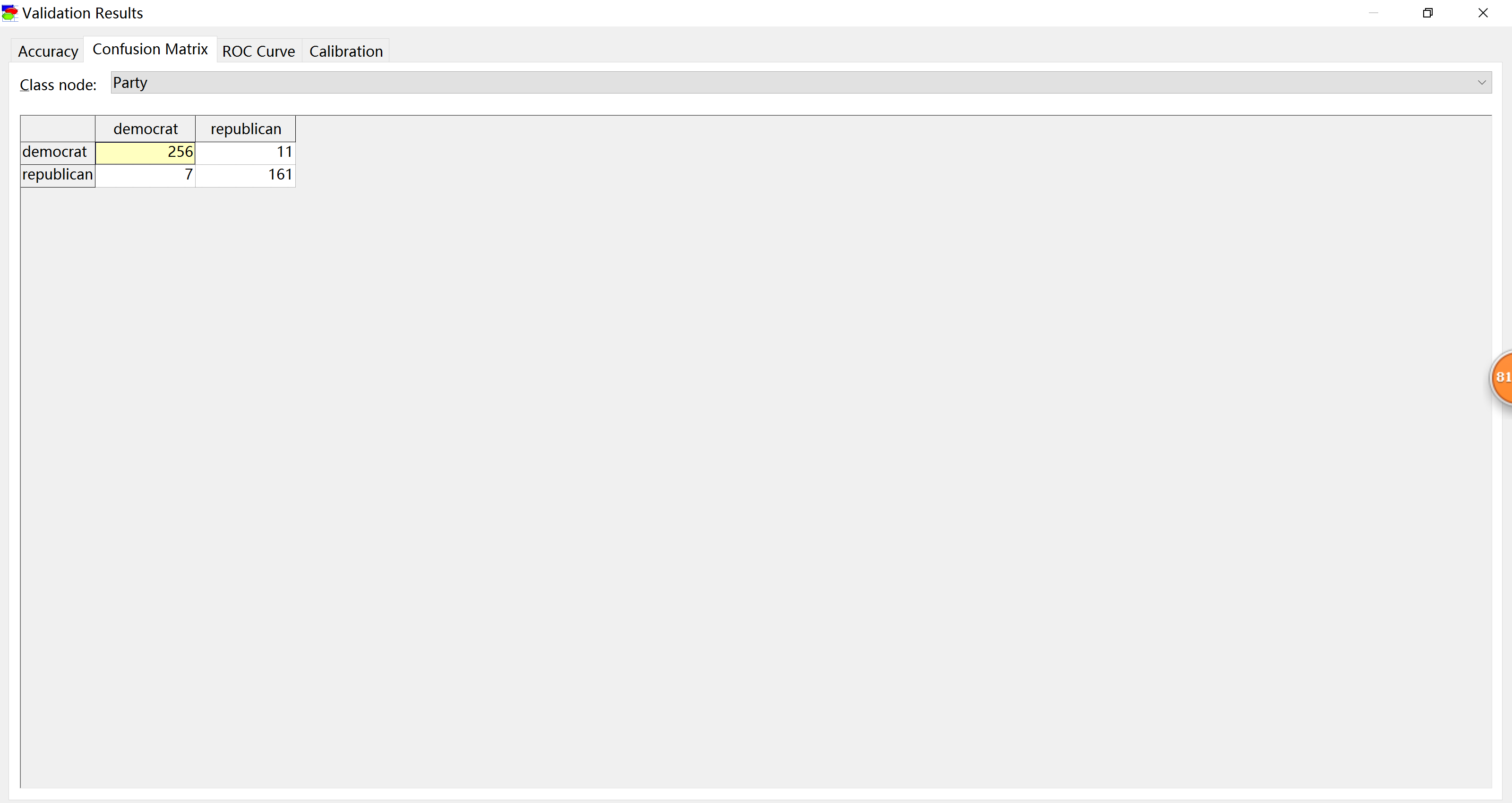




House Votes PC model has 0.958621 of accuracy to predict the actual value.

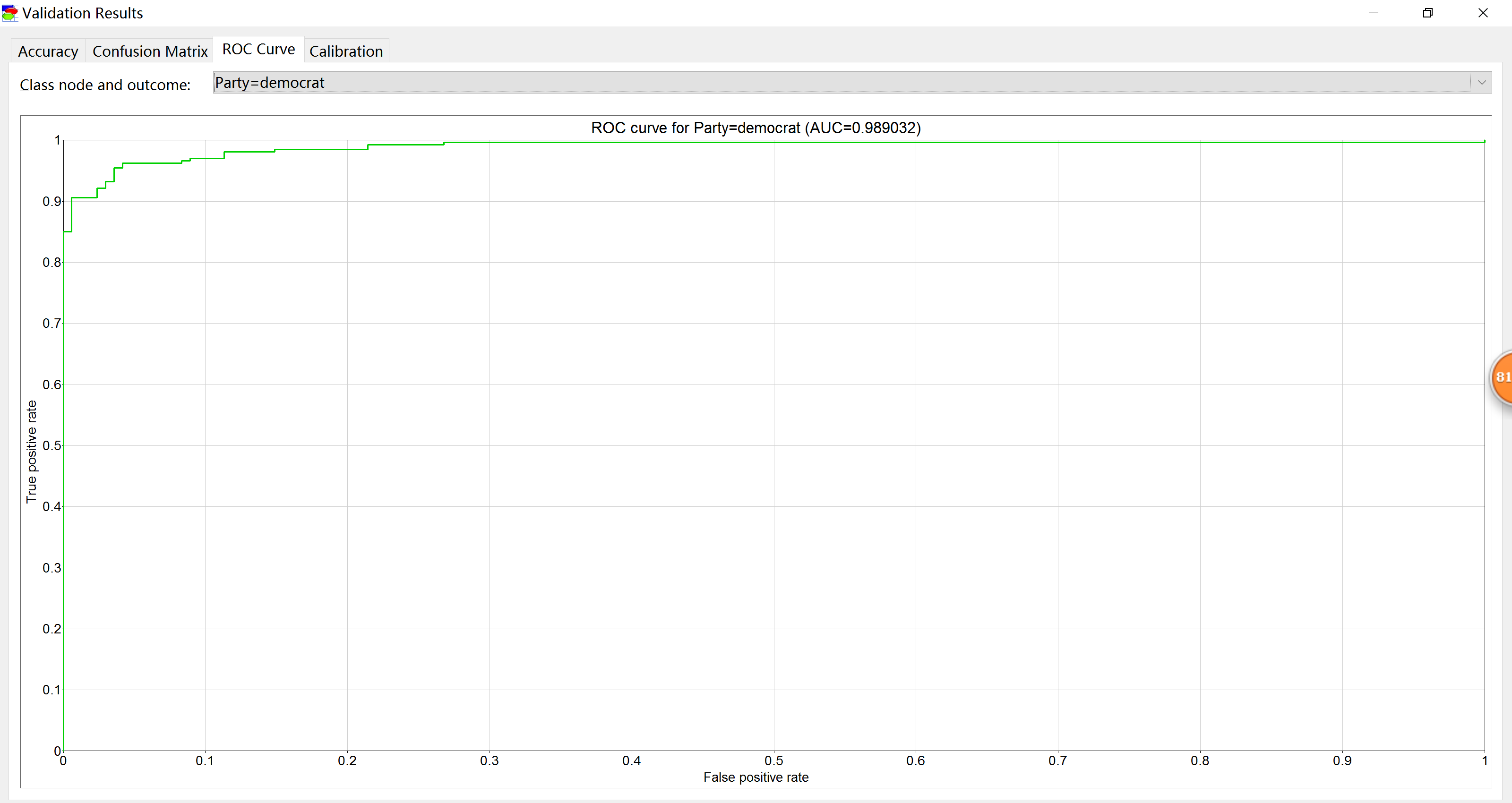


Confusion matrix

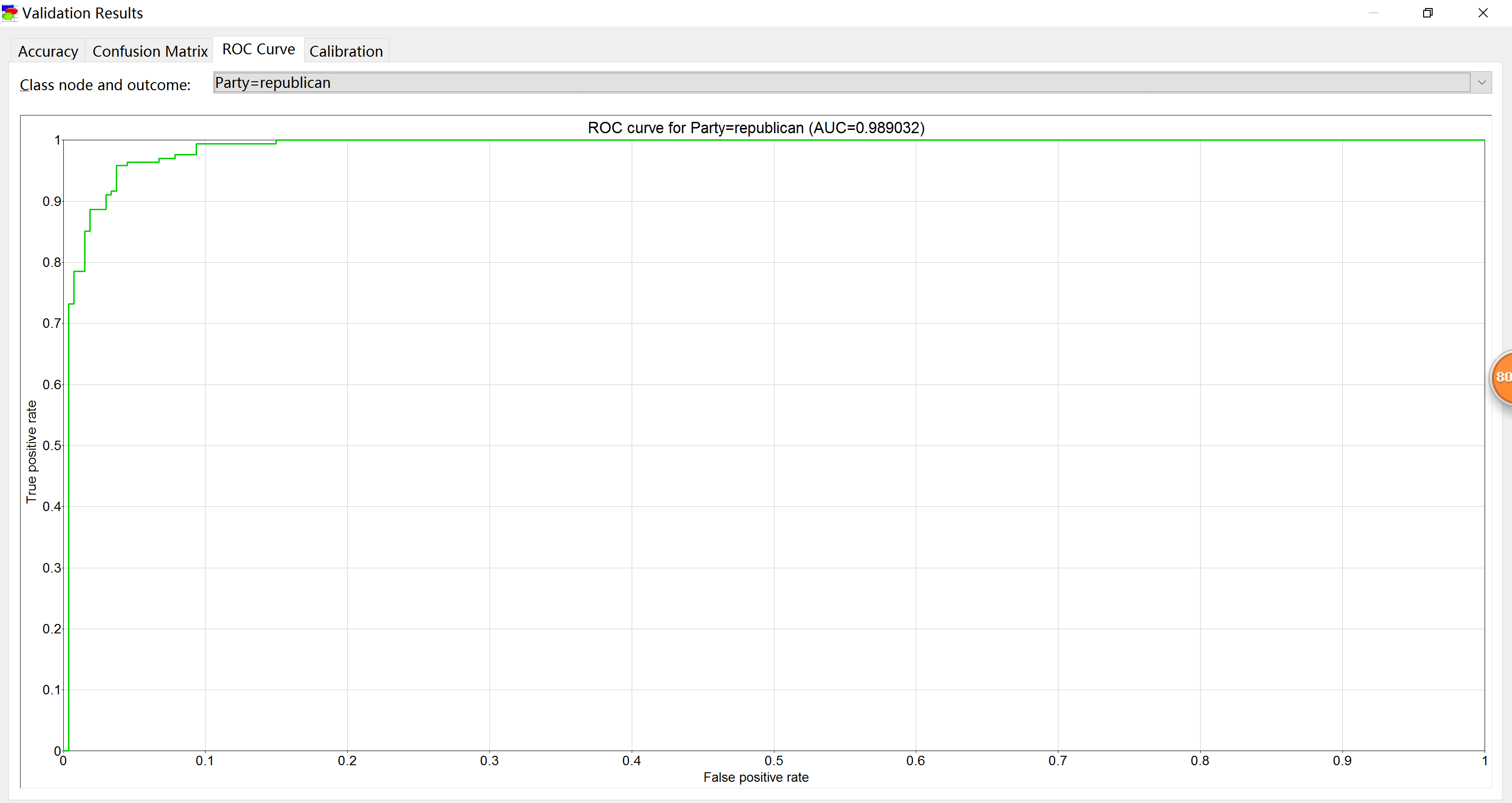


Next, we got the ROC curve for model-3

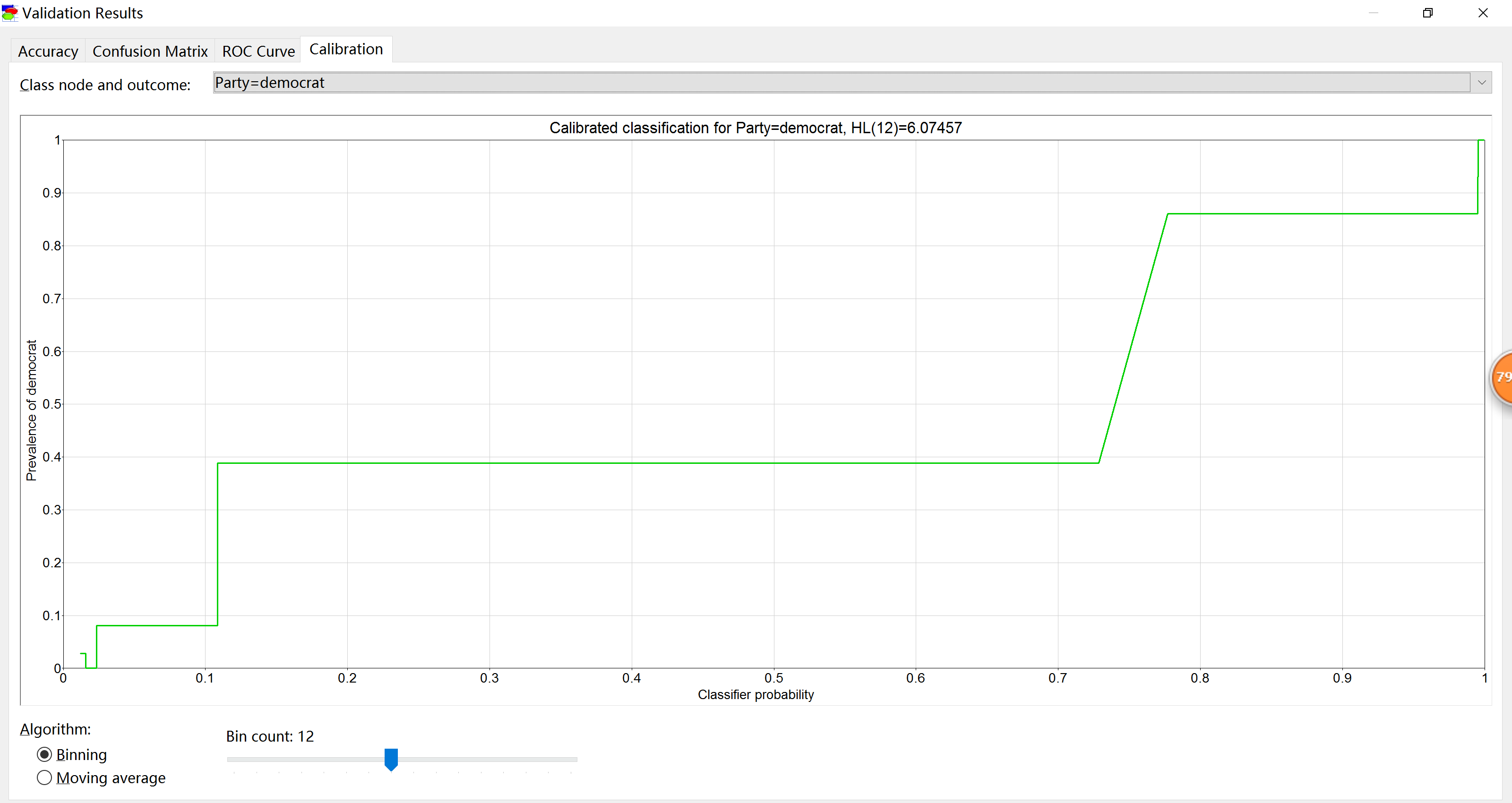
Below is ROC curve with Party = democrat (AUC = 0.989032)

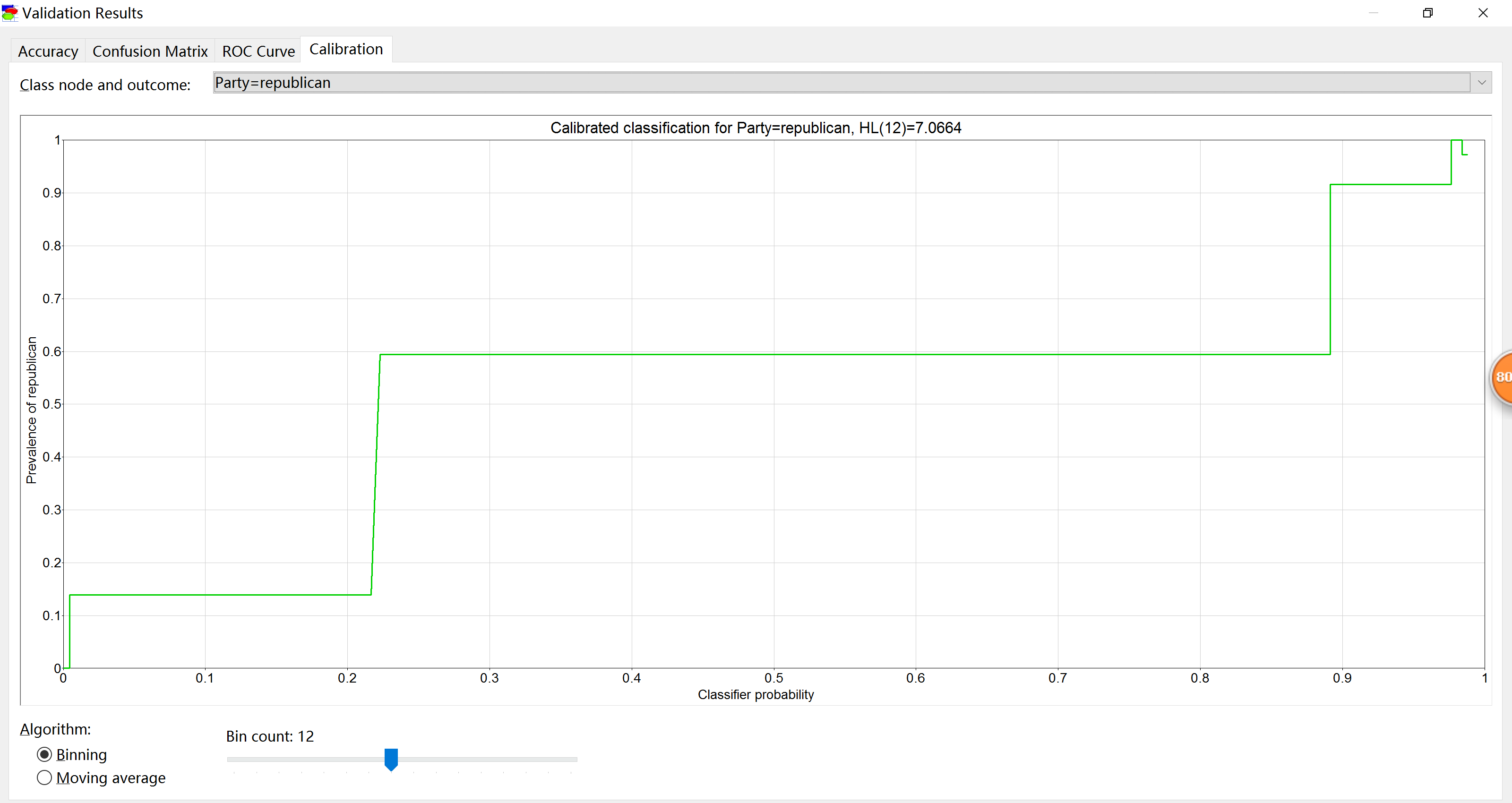


Party = republican



Since we did not change our selected bin count which is count = 12, we generated calibrations.





**Result**

After we compared these three models, we think the first model performs best (with accuracy 0.96092) even though its AUC is not as high as the third model, which means the third model has more accuracy on classification. We still consider general accuracy of performance of the models as a primary factor to decide whether the model is moderately fit in data sets.